# Estimation of Market Share by Using Discretization Technology: An Application in China Mobile

Xiaohang Zhang, Jun Wu, Xuecheng Yang, and Tingjie Lu

Economics and Management School, Beijing University of Posts and Telecommunications,
Beijing China
{zhangxiaohang, junwu, yangxuecheng, lutingjie}@bupt.edu.cn

**Abstract.** The mobile market is becoming more competitive. Mobile operators having been focusing on the market share of high quality customers. In this paper, we propose a new method to help mobile operator to estimate the share in high quality customers market based on the available data, inter-network calling detail records. The core of our method is a discretization algorithm which adopts the Gini criterion as discretization measure and is supervised, global and static. In order to evaluate the model, we use the real life data come from one mobile operator in China mainland. The results prove that our method is effective. And also our method is simple and easy to be incorporated into operation support system to predict periodically.

## 1 Introduction

Due to deregulation, new technologies, and new competitors, the telecommunication industry becomes more competitive than ever. Two biggest mobile operators in China mainland, China Mobile and China Unicom, are struggling for acquiring the customers. As well as attracting the incoming customers, retention of the own high quality customers and acquisition of competitors' customers are becoming the core tasks of each operator. The reason is that these customers are not only the source of the revenue but also the foundation of maintaining an advantage in such a competitive environment. For each operator, the pre-step of retention and acquisition is to estimate the share in high quality customers market and be control of changes of share. This prework is very important because it can help to duly find the potential problems during the process of managing the high quality customers and can be guideline to supervise the operation of sub operators to guarantee that they can be kept in sustaining developments.

In this paper, we propose a new method to help one China Mobile operator in China mainland to estimate the share in high quality customers market. Due to the lack of the foundational data, the estimation can only be based on the inter-network calling detail records (CDRs) which are stored in the operation support systems (OSS). The CDRs describe the calling behaviors between the competitors' customers and the operators' internal customers, which are composed of many fields including calling part, called part, starting time and duration of the call. However, the calling

M. Bubak et al. (Eds.): ICCS 2008, Part II, LNCS 5102, pp. 466–475, 2008.

types, which include local call, long-distance call and roam call, are difficult to identify from the inter-network CDRs. In order to decrease the costs of data process, the estimation can only depend on the total call duration of each customer.

In the traditional method, the estimation process can be decomposed of three steps. Firstly, the distribution of inter-network call duration of China Unicom's customers is computed. Secondly, a cut point is chosen based on the distribution. Finally, the market share can be estimated by using the scale of the customers whose calling duration is greater than the chosen cut point. These customers are considered as high quality ones. The foundational hypothesis of the traditional method is that the customers who have higher inter-network calling duration are more possible to be high quality customers. Although this method is adopted universally by operators, it has some drawbacks and can result in estimation error.

- Due to the difference of the customers' behavioral structure and the difference of the prices of telecom services, the hypothesis of the traditional method can't be satisfied. For example, for some customers, the long-distance calls and roam calls occupy bigger proportion in the total duration than others'. Because the prices of these two services are higher, although they may have shorter total calling duration, they also can be high quality customers.
- The China Unicom customers can give call to many customers including not only China Mobile but also China Telecom and China Netcom ones. However, the OSS of China Mobile only records the CDRs between China Mobile and China Unicom. So if some customers have long calling duration, most of which are not directional to China Mobile, then these customers are possible to be considered as low quality ones.

The most current research of market share estimation focus on the sales prediction [1, 2]. The research on customers' market share estimation in telecom is rare. And because of the disadvantages of traditional method, we propose a new method which is a process of inference from China mobile customers to China Unicom customers. During the process our method adopts supervised discretization technology which has good characters compared with other ones. We estimate the market share based on the real life data which come from one China Mobile operator. And the data are customers' CDRs from January to November in 2007. The results show that our method is simple and comparatively accurate.

## 2 Discretization Method

The discretization methods can be classified according to three axes [3]: supervised versus unsupervised, global versus local, and static versus dynamic. A supervised method would use the classification information during the discretization process, while the unsupervised method would not depend on class information. The popular supervised discretization algorithms contain many categories, such as entropy based algorithms including Ent-MDLP [4, 5], D2 [6], Mantaras distance [7], dependence based algorithms including ChiMerge, Chi2 [8], modified Chi2 [9], Zeta [10], and binning based algorithms including 1R, Marginal Ent. The unsupervised algorithms contain equal width, equal frequency and some other recently proposed algorithms such as PCA-based algorithm [11] and an algorithm using tree-based density estimation [12].

Local methods produce partitions that are applied to localized regions of the instance space. Global methods, such as binning, produce a mesh over the entire continuous instances space, where each feature is partitioned into regions independent of the other attributes.

Many discretization methods require a parameter, $n$, indicating the maximum number of partition intervals in discretizing a feature. Static methods, such as Ent-MDLP, perform the discretization on each feature and determine the value of n for each feature independent of the other features. However, the dynamic methods search through the space of possible $n$ values for all features simultaneously, thereby capturing interdependencies in feature discretization. Most of the popular methods are classified according to the three axes in [13]. The Fig. 1 describes the basic process of discretization. In our application, we use the Gini-criterion based discretization which is supervised, global and static.

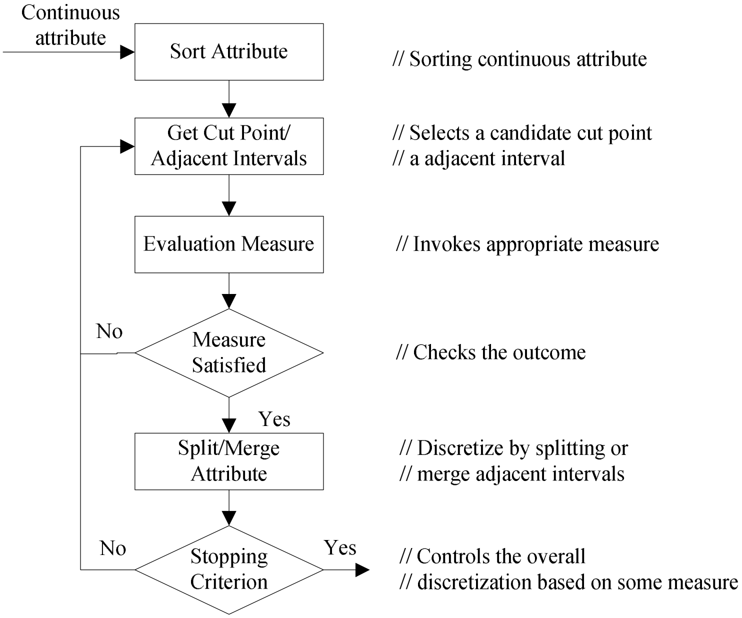
Continuous attribute
Sort Attribute
// Sorting continuous attribute
Get Cut Point/ Adjacent Intervals
// Selects a candidate cut point
// a adjacent interval
Evaluation Measure
// Invokes appropriate measure
No
Measure Satisfied
// Checks the outcome
Yes
Split/Merge Attribute
// Discretize by splitting or
// merge adjacent intervals
No
Stopping Criterion
Yes
// Controls the overall
// discretization based on some measure

**Fig. 1.** Process of discretization

## 2.1 Notations

Suppose for a supervised classification task with $k$ class labels, the training data set $S$ consists of $|S|$ instances, where each instance belongs to only one of $k$ classes. Let $A$ be one continuous attribute. Next, there exists a discretization scheme $D$ on the attribute $A$, which discretizes the attribute $A$ into $n$ discrete intervals bounded by the pairs of numbers:

$$D: \{[b_0, b_1], (b_1, b_2], \cdots (b_{n-1}, b_n]\},$$

where $b_0$ and $b_n$, respectively, are the minimal and the maximal values of the attribute $A$ , and the values in $D$ are arranged in ascending order. The discretization scheme $D$ is called an $n$-scheme. These values constitute the cut points set $B = \{b_0, b_1, \cdots, b_n\}$ of the discretization scheme. The discretization algorithm is to determine the value of $n$ and the cut points set.

### 2.2 Discretization Measure

For the $i$th interval $(b_{i-1}, b_i]$, we can get a conditional class probability $p(i) = (p_1^i, \cdots p_j^i, \cdots p_k^i)$where $p_j^i$ is the $j$th class probability in $i$th interval and satisfies $\sum_{j=1}^{k} p_j^i = 1$. The Gini [14] is defined as follows:

$$\boldsymbol{Gini(interval\, i) = 1 - \sum_{j=1}^{k} \left(p_j^i\right)^2} \qquad (1)$$

In our algorithm we use Gini gain as the discretization measure. The cut point $b$ is chosen based on the criterion, whose Gini gain value is the biggest on attribute $A$. The Gini gain $\Delta G$ is defined as:

$$\Delta \boldsymbol{G(A, b; S) = Gini(S) - \frac{|S_1|}{|S|} Gini(S_1) - \frac{|S_2|}{|S|} Gini(S_2)}, \qquad (2)$$

where $Gini(\cdot)$ is the Gini measure defined in Eq. (1), $S_1$ and $S_2$ are the subsets of $S$ partitioned by the cut point $b$, $|\cdot|$ denotes the number of instances.

### 2.3 Stopping Criterion

The training set is split into two subsets by the cut point which is chosen using Gini measure. Subsequent cut points are selected by recursively applying the same binary discretization method to one of the generated subsets, which has biggest Gini gain value, until the stopping criterion is achieved. Because the quality of discretization methods involves a tradeoff between simplicity and predictive accuracy, the stopping criterion of our algorithm is defined by

$$\boldsymbol{G_{n+1} \ln(n + 1 + p) > G_n \ln(n + p)}, \qquad (3)$$

where $n$ denotes the current number of intervals, $p$ is a positive integer determined by the user, $G_n$is the Gini value with $n$ intervals, defined by

$$\boldsymbol{G_n = \sum_{i=1}^{n} \frac{|S_i|}{|S|} Gini(interval\, i)}. \qquad (4)$$

Eq. (3) can be easily returned as follows:

$$\boldsymbol{G_n / G_{n+1} < ln(n + 1 + p) / \ln(n + p)}. \qquad (5)$$

From the Eq. (5), we can know that the parameter $p$ can affect the number of partition intervals. The smaller the right part of the Eq. (5) is, the more chances the algorithm has to discretize the continuous attribute further. In general, higher $p$ value can result in more intervals.

## 2.4 Comparison with Other Discretization Measures

The following are two other discretization measures.

$$\text{Entropy measure: } \boldsymbol{E(interval\ i)} = -\sum_{j=1}^{k} \boldsymbol{p_j^i \log(p_j^i)} \tag{6}$$

$$\text{Minimal measure: } \boldsymbol{M(interval\ i)} = \mathbf{min}\boldsymbol{(p_1^i, p_2^i, \cdots, p_k^i)} \tag{7}$$

These two measures and Gini measure have common characters. Their values are high for lower probable events and low otherwise. Hence, they are the highest when each event is equi-probable, i.e., $p_j^i = 1/k$ for each $j$; and they are the lowest when $p_j^i = 1$ for one event and 0 for all other events. As known, entropy is one of the most commonly used discretization measures in the discretization literature. When there are only two classes in a classification task, entropy undoubtedly is an excellent measure of class homogeneity. However, when there are more than two classes in a classification problem, entropy sometimes cannot accurately reflect the class homogeneity. For example, see the Table 1.

**Table 1.** Example for comparison of three measures

| Case | P1 | P2 | P3 | Entropy | Gini | Minimal |
|---|---|---|---|---|---|---|
| 1 | 1/2 | 1/2 | 0 | 1 | 0.5 | 0 |
| 2 | 1/6 | 2/3 | 1/6 | 1.25 | 0.5 | 1/6 |
| 3 | 1/8 | 3/4 | 1/8 | 1.06 | 0.406 | 1/8 |

In the example, the entropy value for case 1 is the lowest. Because the smaller value is preferred, the case 1 is the best of all. But the case 3 has higher prediction accuracy, it is 3/4 higher than 1/2 in case 1. Likewise, the minimal measure has the same ordered value as entropy. However, Gini measure gives the case 3 the lowest value. So in this example, Gini measure shows the better ability in terms of classification than the other ones.

We compare the contours of three measures, entropy, Gini, and minimal, in Fig. 2. The vertex of the triangle denotes the event in which only one class label occurs. The center $O = (1/3, 1/3, 1/3)$ denotes that each class label occurs at the equal probability. The farther away the point moves from $O$, the higher the degree of the class heterogeneity.

The minimal measure only consider the minimum class information in an interval, whereas the Gini measure takes into account all the class information and evaluates the interval according to the whole class distribution. The shape of contour of entropy

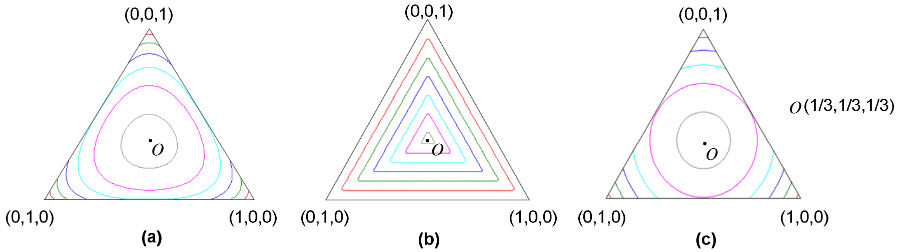

**Fig. 2.** Comparison of the contours of three measures. (a) The contour of entropy. (b) The contour of minimal. (c) The contour of Gini.

lies between minimal and Gini. Entropy and minimal measure seem to prefer the points that are close to the boundary of the triangle. So they prefer (1/2, 1/2, 0) than (1/8, 3/4, 1/8).

# 3 Estimation Method

## 3.1 Notations

All telecom operators pay more attention to average revenue per user (ARPU). The customers with high ARPU mean that they can provide more revenue to operators and they are the high quality customers. The ARPU depends on the calling minutes per user in one month (MOU) and the price of services. In our application, the customers which satisfy the following conditions can be defined as high quality customers in $x$th month.

- The average ARPU of $x\text{th}, x+1\text{th}, x+2\text{th}$ months is greater than 100 Yuan.
- During $x\text{th}, x+1\text{th}, x+2\text{th}$ months, the calling behaviors all exist.

In order to give the basic definitions, the customers' average MOU of $x\text{th}, x+1\text{th}, x+2\text{th}$ months are sorted by ascending order. Then the MOU values are discretized into many intervals by cut points. The number of customers and high quality customers falling into these intervals can be computed. The Fig. 3 shows the process.

In Fig. 3, $cutP_i (i = 1,2,\cdots,n)$ represents the $i$th cut point, $M_i$ is the number of the China Mobile customers falling into the $i$th MOU interval, $MIP_i$ is number of the high

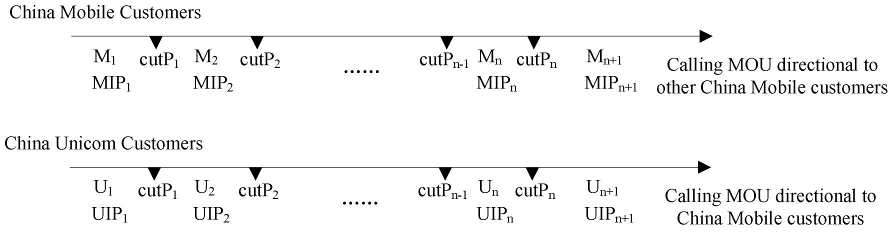

**Fig. 3.** Discretization of MOU

quality customers falling into $i$th MOU interval. And $U_i$ is the number of the China Unicom customers falling into the $i$th MOU interval, $UIP_i$ is number of the high quality customers falling into $i$th MOU interval. MOU means the average MOU of the continuous three months.

### 3.2 Basic Hypothesis

The basic hypothesis of our method is that in all discretized intervals, proportions of the high quality customers for two operators are similar, which can be represented by

$$\frac{\boldsymbol{MIP_i}}{\boldsymbol{M_i}} = \frac{\boldsymbol{UIP_i}}{\boldsymbol{U_i}}, \boldsymbol{i = 1, 2, \cdots, n, n + 1.} \tag{8}$$

The meaning of the hypothesis is concluded in the following.

- In each MOU interval, the distribution of the proportion between the China Unicom customers' inter-network call MOU and their total call duration is similar to the China Mobiles' customers.
- For China Mobile and China Unicom customers, the call behavior structure is similar in each MOU interval.

Because of the similarity of service structure, service quality and service price for the two operators, the hypothesis can be approximately satisfied. And to some extent, the hypothesis depends on the chosen cut points which discretize the MOU into intervals. We adopt the Gini discretization method to choose the cut points.

### 3.3 Estimation

Based on the basic hypothesis, the formula of estimating the number of China Unicom's high quality customers is as follows:

$$\boldsymbol{UIP} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n+1}} \boldsymbol{UIP_i} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n+1}} \frac{\boldsymbol{MIP_i}}{\boldsymbol{M_i}} \boldsymbol{U_i}, \tag{9}$$

where $MIP_i$ and $M_i$ can be computed simply based on internal CDRs and $U_i$ can be computed based on inter-network CDRs.

In order to estimate the market share of high quality customers for each sub mobile operators, we also need to estimate the number of high quality customers of each sub China Unicom operator.

$$\boldsymbol{UIP^j} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n+1}} \boldsymbol{UIP_i^j} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n+1}} \frac{\boldsymbol{MIP_i^j}}{\boldsymbol{M_i^j}} \boldsymbol{U_i^j}, \tag{10}$$

where $UIP^j$ represents the number of high quality customer in $j$th sub China Unicom operator, $MIP_i{}^j$ represents the number of high quality customers who fall into $i$th MOU interval in $j$th sub China Mobile operator, $M_i{}^j$ is the number of customers falling into $i$th MOU interval in $j$th sub China Mobile operator, and $U_i{}^j$ is the number of customers who fall into $i$th MOU interval in $j$th sub China Unicom operator.

### 3.4 Model Evaluation

Due to lack of the real market share data, we can't directly evaluate the model. We adopted the self-validation method. In other words, we use our method to estimate the proportion of high quality customers to all customers in China Mobile which can be compared to the real proportion. We use the data from *n*th to *n*+2th months to build model and estimate the proportion of *n*+*3*th month. The process is shown in Fig. 4.

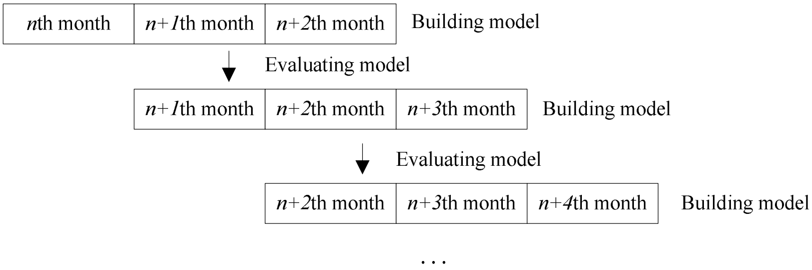

**Fig. 4.** Process of evaluation of model

## 4 Results

In the application, we extract the internal CDRs and inter-network CDRs from January to November in 2007. The cut points obtained by using supervised discretization technology are shown in Table 2. Here the best number of cut points is seven, so there are eight discretization intervals. Because of totally eleven months, we can get nine groups of cut points. In order to get robust results, we adopt the median of all groups as the final cut points.

**Table 2.** The cut points obtain by using supervised discretization method

| cutP | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | Median |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 64 | 70 | 74 | 78 | 80 | 90 | 84 | 87 | 90 | 80 |
| 2 | 115 | 120 | 126 | 132 | 123 | 143 | 141 | 144 | 140 | 132 |
| 3 | 160 | 165 | 165 | 182 | 173 | 191 | 194 | 199 | 190 | 182 |
| 4 | 210 | 217 | 208 | 229 | 224 | 242 | 255 | 246 | 245 | 229 |
| 5 | 268 | 264 | 254 | 277 | 284 | 302 | 326 | 301 | 312 | 284 |
| 6 | 318 | 338 | 341 | 367 | 353 | 383 | 407 | 384 | 400 | 367 |
| 7 | 440 | 436 | 465 | 527 | 476 | 501 | 524 | 512 | 481 | 481 |

The evaluation results are shown in Table 3, in which Ci (i=1, 2, …, 12) represents the *i*th sub China Mobile operator. Totally there are twelve sub operators. And because of eleven months, we can obtain eight groups of errors. We can see that the average error of each sub operator is less than 6%, which proves that our estimation method can obtain good accuracy. The Table 4 shows the results of market share estimation. From the description, we can see that our estimation method is so simple that it can be easily incorporated into operation support system to predict periodically.

**Table 3.** The estimation errors

| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | AVE | STD |
|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 2.0% | 2.7% | 3.6% | 2.5% | 1.6% | 2.1% | 3.5% | 3.4% | 2.7% | 0.7% |
| C2 | 3.7% | 10.5% | 6.5% | 4.4% | 3.1% | 3.5% | 7.1% | 6.6% | 5.7% | 2.5% |
| C3 | 3.7% | 3.7% | 4.2% | 3.1% | 2.2% | 1.7% | 2.3% | 2.8% | 3.0% | 0.9% |
| C4 | -7.6% | 3.9% | 3.1% | 2.1% | 1.9% | 1.0% | 2.3% | 3.3% | 1.3% | 3.7% |
| C5 | 6.2% | 8.5% | 10.2% | 7.7% | 3.4% | 3.8% | 3.4% | 2.6% | 5.7% | 2.8% |
| C6 | 4.8% | 5.0% | 6.0% | 3.0% | 2.1% | 0.4% | 1.8% | 1.7% | 3.1% | 2.0% |
| C7 | 3.5% | 2.7% | 5.4% | 2.0% | 0.1% | 0.3% | 1.9% | 0.7% | 2.0% | 1.8% |
| C8 | 6.1% | 3.8% | 4.6% | 4.2% | 4.3% | 2.6% | 2.1% | -0.3% | 3.4% | 1.9% |
| C9 | 3.5% | 1.8% | 4.3% | 4.0% | 3.3% | 1.9% | 3.4% | 3.8% | 3.3% | 0.9% |
| C10 | 3.9% | 6.3% | 7.7% | 5.1 | 3.7% | 1.9% | 1.4% | 1.4% | 3.9% | 2.3% |
| C11 | 5.1% | 6.3% | 8.5% | 6.2% | 4.9% | 4.4% | 4.5% | 5.5% | 5.7% | 1.3% |
| C12 | 1.6% | 5.9% | 4.8% | 3.2% | 0.6% | 1.1% | 1.6% | 1.8% | 2.6% | 1.9% |
| AVE | 3.0% | 5.1% | 5.7% | 3.9% | 2.6% | 2.1% | 3.0% | 2.8% | | |

**Table 4.** Estimation of share in high quality customers market

| | Mar | Apr | May | Jun | July | Aug | Sep | Oct | Nov |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 83.92% | 84.12% | 84.52% | 84.82% | 84.97% | 85.12% | 85.49% | 86.22% | 86.93% |
| C2 | 74.86% | 75.10% | 76.71% | 77.41% | 77.62% | 77.54% | 77.86% | 78.75% | 79.67% |
| C3 | 78.54% | 78.89% | 79.16% | 79.59% | 80.00% | 80.43% | 80.80% | 81.56% | 82.31% |
| C4 | 80.06% | 78.39% | 78.63% | 79.08% | 79.58% | 80.10% | 80.62% | 81.46% | 82.23% |
| C5 | 75.28% | 75.83% | 77.21% | 78.54% | 79.20% | 79.72% | 80.34% | 81.30% | 82.02% |
| C6 | 86.30% | 86.62% | 86.91% | 87.36% | 87.78% | 88.03% | 88.20% | 88.37% | 88.67% |
| C7 | 87.89% | 88.02% | 87.85% | 88.07% | 88.05% | 88.03% | 88.09% | 88.47% | 88.90% |
| C8 | 79.32% | 79.54% | 80.02% | 80.74% | 81.44% | 81.92% | 82.29% | 82.78% | 83.34% |
| C9 | 76.65% | 77.15% | 77.41% | 78.01% | 78.45% | 78.65% | 79.14% | 79.74% | 80.64% |
| C10 | 86.51% | 86.65% | 87.11% | 87.72% | 88.36% | 88.78% | 89.33% | 89.91% | 90.52% |
| C11 | 82.96% | 83.29% | 83.88% | 84.66% | 85.40% | 86.06% | 86.62% | 87.33% | 87.92% |
| C12 | 86.38% | 86.33% | 86.99% | 87.53% | 87.82% | 87.87% | 88.31% | 88.99% | 89.52% |

## 5 Conclusions

In this paper, we propose a new method to estimate the market share of high quality customers for mobile operators. This method is based on a discretization technology which adopts the Gini criterion as discretization measure. The Gini-based discretization method is supervised, static and global, which is compared with other methods and has its' good characters. We describe the complete process of estimating market share. And based on real life data come from one China mobile operator, the estimation method is implemented. The results prove that our method is effective. Our method is also simple and can be easily incorporated into the OSS to predict periodically.

## References

1. Kumar, V., Anish, N., Rajkumar, V.: Forecasting category sales and market share for wireless telephone subscribers: a combined approach. International Journal of Forecasting 18(4), 583–603 (2002)
2. Fok, D., Franses, P.H.: Forecasting market shares from models for sales. International Journal of Forecasting 17(1), 121–128 (2001)
3. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. In: Proc. 12th Int'l Conf. Machine Learning, pp. 194–202 (1995)
4. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. Thirteenth International Joint Conference on Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann, San Francisco (1993)
5. Fayyad, U., Irani, K.: Discretizing continuous attributes while learning bayesian networks. In: Proc. Thirteenth International Conference on Machine Learning, pp. 157–165. Morgan Kaufmann, San Francisco (1996)
6. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS, vol. 482, pp. 164–177. Springer, Heidelberg (1991)
7. Cerquides, J., Mantaras, R.L.: Proposal and empirical comparison of a parallelizable distance-based discretization method. In: KDD 1997: Third International Conference on Knowledge Discovery and Data Mining, pp. 139–142 (1997)
8. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Vassilopoulos, J.F. (ed.) Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, Herndon, Virginia, November 5-8, 1995, vol. 1995, pp. 388–391. IEEE Computer Society Press, Los Alamitos (1995)
9. Tay, F.E.H., Shen, L.X.: A Modified Chi2 Algorithm for Discretization. IEEE Trans. Knowledge and Data Eng. 14(3), 666–670 (2002)
10. Ho, K.M., Scott, P.D.: Zeta: A global method for discretization of continuous variables. In: KDD 1997: 3rd International Conference of Knowledge Discovery and Data Mining. Newport Beach, CA, pp. 191–194 (1997)
11. Sameep, M., Srinivasan, P., Hui, Y.: Toward Unsupervised Correlation Preserving Discretization. IEEE Transaction on Knowledge and Data Engineering 17(8) (August 2005)
12. Gbi, S., Eibe, F.: Unsupervised Discretization using Tree-based Density Estimation. Lecture Notes in Computer Science (2006)
13. Huan, L., Farhad, H., Lim, T.C., Manoranjan, D.: Discretization: An Enabling Technique. Data Mining and Knowledge Discovery 6, 393–423 (2002)
14. Leo, B., Jerome, F., Charles, J.S., Olshen, R.A.: Classification and Regression Trees. Wadsworth International Group (1984)
15. Xiao-Hang, Z., Jun, W., Ting-Jie, L., Yuan, J.: A Discretization Algorithm Based on Gini Criterion. In: Machine Learning and Cybernetics, 2007 International Conference, August 19-22, 2007, vol. 5, pp. 2557–2561 (2007)